3D Visual Passcode: Speech-driven 3D Facial Dynamics for Behaviometrics

3D Visual Passcode: Speech-driven 3D Facial Dynamics for Behaviometrics

Jie Zhang*, Robert B. Fisher†

January 31, 2019

Abstract

Face biometrics have achieved remarkable performance over the past decades, but unexpected spoofing of the static faces poses a threat to information security. There is an increasing demand for stable and discriminative biological modalities which are hard to be mimicked and deceived. Speech-driven 3D facial motion is a distinctive and measurable behavior-signature that is promising for biometrics. In this paper, we propose a novel 3D behaviometrics framework based on a "3D visual passcode" derived from speech-driven 3D facial dynamics. The 3D facial dynamics are jointly represented by 3D-keypoint-based measurements and 3D shape patch features, extracted from both static and speech-driven dynamic regions. An ensemble of subject-specific classifiers are then trained over selected discriminative features, which allows for a discriminant speech-driven 3D facial dynamics representation. We construct the first publicly available Speech-driven 3D Facial Motion dataset (S3DFM) that includes 2D-3D face video plus audio samples from 77 participants. The experimental results on the S3DFM show that the proposed pipeline achieves a face identification rate of 96.1%. Detailed discussions are presented, concerning anti-spoofing, head pose variation, video frame rate, and applicability cases. We also give comparison with other baselines on "deep" and "shallow" 2D face features.

*Jie Zhang was with Beihang University, China, and the University of Edinburgh, UK. (zhangjie09@buaa.edu.cn)

†Robert B. Fisher is with the School of Informatics, the University of Edinburgh, EH8 9BT, UK. (rbf@inf.ed.ac.uk)

1 Introduction

Human faces are a discriminative biological representation for identities. Thus, face biometrics have been widely applied in access control and information security. Compared with traditional token or knowledge based information protection, human faces are impossible to be forgotten. Compared with other bio-modalities such as fingerprints, palmprints, and DNA, face biometrics are non-contact and non-invasive.

Over the past decades, 2D face biometrics have had great success, which benefits from rich 2D image representations [1, 2]. However, common weaknesses of 2D face biometrics are high sensitivity to the application conditions, e.g. illumination, shadows, facial pose, and appearance (makeup, complexion, aging, etc). With the popularity of various 3D sensors, 3D face biometrics have gained increasing attention. The most prominent advantages of 3D data are immunity to texture-related variations and pose-invariance, which also allow for fine-level recognition. Apart from the data modality, anti-spoofing is another significant property of biometrics algorithms. Due to the openness of social networking, static face data is easily intercepted and misused, which degrades the reliability of biometric systems. Since 3D facial dynamics contain extra informative and discriminative power for characterizing individuals, it is a promising bio-modality for enhancing information security.

In this paper, we present person recognition with a 3D individual-specific bio-modality, which is less likely to be deceived or mimicked and is robust against head pose variations. We propose a novel behaviometric method using speech-driven 3D facial dynamics as a "3D visual passcode". The core merits of the "3D visual passcode" are two-fold. Firstly, the 3D bio-modality is a cooperative and text-constrained behavior generated from a human speaking a real passcode. In comparison to spontaneous facial expressions, it is more repeatable and possesses the inherent advantages of 3D data. Secondly, compared with static faces, it is relatively tough to be mimicked and deceived due to being person specific and distinctive. The proposed idea can be generally applied in any biometric system where 3D video scanners are installed facing users. Also, our framework and its speaking-dynamics features can be generalized to any spoken passcodes (set by users) and is invariant to speaking speed, which in turn means that the uniqueness of the "3D visual passcode" originates from both the subject-specific facial motion and the privacy of a passcode.

The properties mentioned above of the proposed "3D visual passcode" raise 3 main sub-problems that will be investigated and discussed in this paper. (1) Repeatability: when a subject speaks the same passcode at different times, the "3D visual passcode" should be consistent. (2) Distinctiveness: even if different subjects speak the same passcode, the "3D visual passcode" should be distinctive or subject-specific. (3) Anti-spoofing: the "3D visual passcode" should be difficult to be mimicked and fabricated by others.

The contributions of the paper are:

• A behaviometrics algorithm using 3D speech-driven dynamic faces (Section 4.1).

• The first publicly available Speech-driven 3D Facial Motion dataset – S3DFM[1] (Section 3). The dataset includes of over 1000 samples from 77 participants. Each sample consists of a one second 3D point cloud sequence plus a pixel-wise registered 2D intensity sequence captured at 500 frames per second, and a synchronized audio sequence. Note that the proposed dataset is the largest known dataset based on a speech-driven 3D facial motion. The new dataset can be generally used for various speech-driven recognition research, such as identity recognition, gender/age estimation, liveness detection, etc.

The experiments (Section 5) demonstrate the power of our approach. We even made the task more difficult by asking all the participants to use the same passcode, and still achieved 740 correct identifications out of 770 trials (96.1%) with 77 participants. A few important abbreviations are listed in Table 1 below.

Table 1: Abbreviations of a few important items

Item	**Abbr.**	**Item**	**Abbr.**
Facial Landmark	FLM	Cumulative Match Characteristic	CMC
Principal Curvature	PC	Deep Neural Network	DNN
Speech-driven 3D Facial Motion Dataset			S3DFM

2 Related works

This section briefly reviews relevant work on 3D face behaviometrics and 3D dynamic face datasets. We know there are numerous topics related with

[1]http://groups.inf.ed.ac.uk/trimbot2020/DYNAMICFACES/

our work, but here we only focus on the closest topics below. To the best of our knowledge, the new dataset (S3DFM) is the first publicly available speech-driven 3D dynamic face dataset in the community.

Table 2: Summary of existing 3D dynamic face datasets

Dataset	Motion Type	Sequences	Subjects	Sensor / Frame Rate (fps)	Public	Year
Chang et al. [3]	6 expressions	36	6	NTSC camera/projector sensor / 30	N	2005
BU-4DFE [4]	6 expressions	606	101	DI3D triple cameras / 25	Y	2008
BP4D-Spontaneous [5]	8 expressions	-	41	DI3D triple cameras / 25	Y	2013
VT-KFER [6]	6 expressions	1988	32	Kinect 1.0	Y	2015
D3DFACS [6]	19 single AUs + 97 combined AUs	519	10	3DMD 3D stereo sensor / 60	Y	2011
Benedikt et al. [7]	expressions, AUs, reading 9 words	606	94	3DMD Face Dynamic System / 48	N	2010
Hi4D-ADSIP [8]	14 articulations	3360	80	DI3D triple cameras / 25	Y	2011
Ours: S3DFM	dynamics: repeated one passcode 10 times; dual-dynamics: speaking and head moving	770 + 260	77 + 26	DI4D stereo video sensor / 500	Y	2019

2.1 From 3D face biometrics to behaviometrics

2D face recognition algorithms are usually based on rich texture information. A lot of newly developed texture and image representations [1, 2, 9] are promising for the task of face recognition. 3D face recognition benefits from using real 3D geometric information that underpins the properties of pose invariance and illumination invariance, etc. Thus, 3D face recognition is an active topic in biometrics. The majority of existing approaches are based on hand-crafted features [10, 11] or 3D Morphable Model (3DMM) fitting [12, 13]. The low-level approaches based on hand-crafted features have explicable descriptiveness and are powerful enough to handle normal scales of data, but they usually depend on algorithmic operations with a high complexity. The 3DMM-based approaches use parametric face representation but also suffer from a high computation cost from model fitting and optimization. Recently, data-driven-based approaches via end-to-end learning models have been used for 3D face biometrics. Kim et al. [14] proposed a 3D face biometrics approach based on transfer learning, which utilizes a convolutional neural network (CNN) pre-trained on 2D intensity images to induce a fine-tuned CNN model specialized for 3D data representation. The method avoids training a CNN from scratch using a large dataset of 3D facial scans. All of the above approaches are based on single face scans, without dynamic information.

Neurophysiological research demonstrates that dynamic information enhances visual perception by conveying more discriminative information [15]. The research establishes a theoretical basis for face behaviometrics [16], including identity, gender, age, and ethnicity estimation. Here, we separate facial dynamics as speech-driven and non-speech related. Earlier works, e.g. [17] extracted dynamic visual features from 2D lip motions to enhance access security. [18, 19] further combine 2D lip motion from an utterance with audio information for speaker recognition. For non-speech related facial dynamics, Zafeiriou et al. [20] proposed that facial deformation in spontaneous smile/laughter is useful in behaviometrics, although it is difficult to produce genuine expressions on demand. Dantcheva et al. [21] investigated how dynamic features from smiles encrypt gender evidence and proposed Smile-Dynamics for gender estimation. All of the above behaviometric methods are based on 2D intensity sequences without 3D geometry or shape information. For 3D face behaviometrics, Benedikt et al. [22, 7] compared the uniqueness and permanence of 3D dynamic faces performing short verbal or nonverbal motions (facial expressions), and concluded that verbal motions are more repeatable and reliable for biometrics. They thus proposed a face behaviometrics algorithm that quantizes 3D facial motions using dynamic eigen-coefficients of a PCA-based 3D face morphable model, with a matching algorithm based on weighted dynamic time warping (WDTW). In summary, 3D face behaviormetrics is a promising yet under-explored topic.

2.2 3D facial dynamics representation

Most existing 3D facial dynamics research focuses on 3D dynamic facial expressions [23]. The pioneers of investigating facial dynamics extracted facial action units (AU) of six spontaneous expressions and established the well-known Facial Action Coding System (FACS) [24]. The 3D facial dynamics representation can be categorized into point-based and part-based methods. The point-based methods mainly contain keypoint-tracking-based local dynamics representations [25, 26] and dense-point-based global dynamics representations, e.g. Facial Level Curve [27], Free-Form Deformation [28], LBP-TOP on 3D flow matrices [29], Dense Scalar Field on radial curves [30]. The part-based methods contain time-varying 3D shape index descriptors [31], curvature-based spatio-temporal 4D facial features - Nebula Feature [32], 3D discrete cosine transform based spatio-temporal features [33], and ST-GeoTopo+ descriptors [34]. Although both facial expression and identity

recognition focus on 3D dynamic face information, the former investigates the discriminability of different non-verbal motions, while the latter focuses on the discriminability and repeatability of different subjects in one fixed 3D facial motion class.

2.3 3D dynamic face datasets

Existing 3D dynamic face datasets include 3D dynamic facial expression datasets (spontaneous/non-spontaneous) [3, 4, 5, 6], 3D facial AU datasets [35], and comprehensive 3D facial motion datasets [7, 8]. Detailed properties can be seen in Table 2. The first 3D dynamic facial expression dataset [3] only contains 6 subjects performing 6 basic expressions, without public availability. The most popular 3D dynamic face dataset is BU-4DFE dataset [4] (based on the BU-3DFE dataset [36]) from Binghamton University. The VT-KFER [6] dataset collects RGBD+time data of 7 annotated expressions, captured by Kinect sensors. In terms of AU datasets, D3DFACS [35] focuses on facial AUs for dynamic morphable facial modeling or expression recognition. BP4D-Spontaneous also utilizes FACS to label frame-level ground truth of facial AUs. All of the above 3D dynamic face datasets only include 3D facial expressions. Hi4D-ADSIP dataset [8] is a high-resolution 3D dynamic facial articulation database used for both expression recognition and clinical diagnosis of facial dysfunctions. The speech-driven dynamic face samples are not enough for the behaviometrics proposed here.

In addition, our new dataset also has a relation with audio-visual (AV) dual-modality dataset. Existing AV datasets are collected at various scenes, e.g. meetings, long-time network TV [37], indoor or outdoor scenes with varying lighting [38]. There are also some audio-visual corpora with sentences or dialogs, such as VidTIMIT [39], AusTalk [40], MOBIO [41]. However, most of the visual modality are based on 2D intensity data. The IEMOCAP dataset [42] involves 3D facial landmarks, but the landmarks were obtained via physical markers attached to the faces of participants. Additionally, the IEMOCAP dataset focuses on emotions elicited from sessions instead of speaking behavior itself.

3 New dataset: Speech-driven 3D Facial Motion Dataset (S3DFM)

To the best of our knowledge, there is no publicly available dataset that focuses on speech-driven 3D facial dynamics across different subjects. We constructed the first public one – S3DFM that would contribute to research on behaviometrics, gender estimation, age estimation, etc. The dataset also includes visual-audio data, where the participant is speaking with his/her head moving randomly. Additionally, we provide high temporal resolution and 10-fold repeatable samples, which are not available in existing datasets. The dataset can be found at:

http://groups.inf.ed.ac.uk/trimbot2020/DYNAMICFACES

The S3DFM dataset contains 2 parts: Frontal Pose (FP) and Varying Pose (VP).There are 1030 sets of samples (515,000 images + 3D models) in total. Each set of sample consists of a 2D dynamic face video, a 3D dynamic face video, and a synchronized audio stream. In the FP part, there are 770 sets of samples from 77 participants, i.e. 10 sets of samples from each of 77 participants. The current 77 participants are from more than 20 nationalities, different ages, genders, ethnicities, etc. In detail, there are 50 males and 27 females. 31 participants are native-Chinese speakers and the rest (46 subjects) are non-native Chinese speakers. The ages range from 16 to 73 years old. The majority of the participants are students and staff from the School of Informatics at the University of Edinburgh. In the VP part, there are 26 participants. Each participant also provides 10 2D-3D dynamic face sequences with synchronized audio sequences. More information is detailed below.

3.1 Data acquisition

The data acquisition device is a high frame rate 3D video sensor (500 fps) from DI4D Ltd [43], as shown in Fig.1. The sensor is a binocular stereo vision system that mainly consists of two infrared intensity cameras. In the acquisition of frontal face data (FP), each participant was asked to spontaneously repeat a short passcode − "ni'hao" (Chinese for "hello") 10 times in front of the sensor, with the head naturally looking straight at the sensor. In the acquisition of varying head pose data (VP), each participant spontaneously repeated the same passcode while moving his/her head randomly.

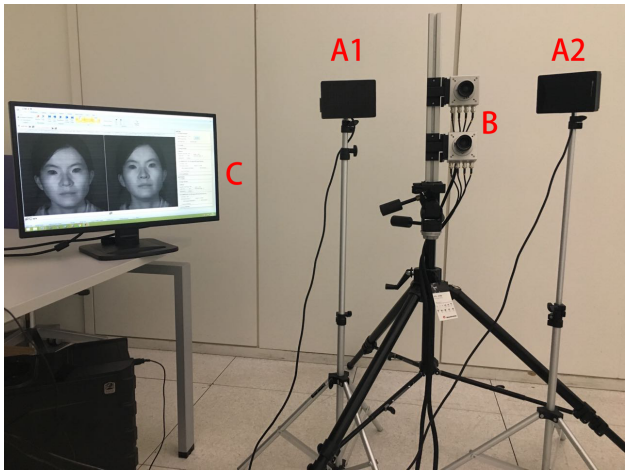

Figure 1: Setup for capturing dynamic face data: A1&A2 are infrared sources to improve brightness; B is a stereo video sensor from DI4D Ltd; C is the data acquisition and reconstruction software.

For each trial, we captured a pair of 2D intensity sequences using the 3D video sensor and a synchronized audio sequence using a microphone. The synchronized audio was not used in the validation algorithm here, but it was used in another 3D video-audio recognition related research [44].

Note that although we used the same passcode across all the trials, the algorithm we proposed below is not specific to the passcode "ni'hao" and can be generally applied to the facial dynamics when speaking any passcodes. In real applications, a client would choose a private passcode. The participant recruitment and data processing were conducting over 1 year, as we prefer to include participants with diverse characteristics. Additionally, since the video and audio acquisition devices are not integrated, manual synchronization is a time-consuming procedure as well.

3.2 Data processing

In the postprocessing, 3D point cloud sequences were reconstructed from the pairwise 2D intensity sequences using DI4D's commercial software with additional spatial smoothing and temporal filtering. Video-audio synchronization was achieved using a camera flash that can be "seen" by the cameras and be "heard" by the microphone simultaneously. Finally, each set of data consists

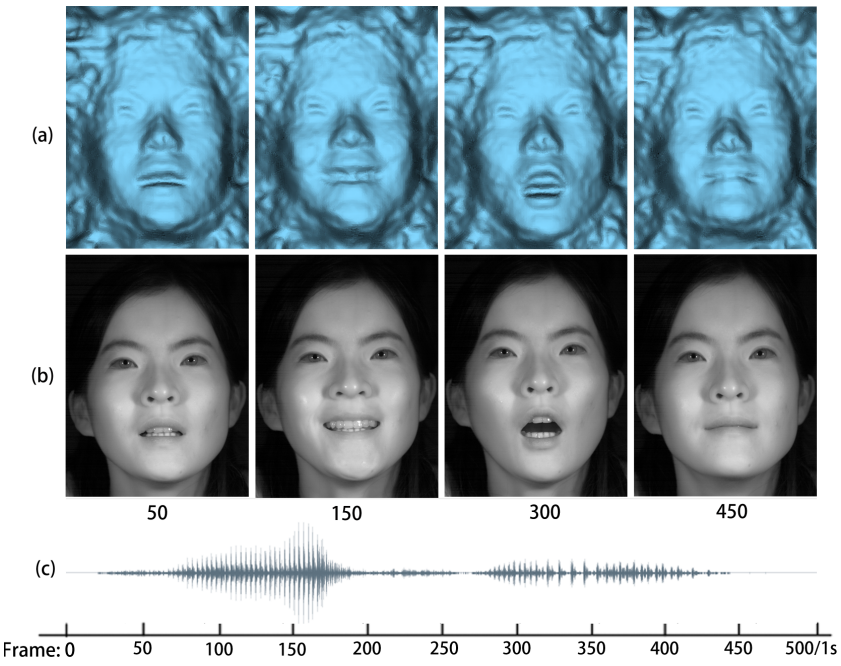

Figure 2: (a) 4 representative 3D frames in a 3D sequence; (b) pixel-wise registered 2D intensity frames; (c) synchronized audio sequence.

of a 3D video and a pixel-wise registered intensity video, plus a synchronized 16kHz audio "passcode". The 1 second video sequence contains 500 frames. The resolutions of the 3D and intensity frames are 600×600 points and 600×600 pixels, respectively. (The 3D frames were downsampled from the original resolution of 1200×1200 pixels in order to improve the processing efficiency and to reduce 3D noise). Example data is shown in Fig.2.

4 Proposed behaviometrics

4.1 Overview

The proposed system framework has two main parts, as illustrated in Fig.3: (1) speech-driven 3D facial dynamics representation; (2) offline database training and online behaviometrics. We overview the proposed method here and give more details in Section 4.2.

For a participant who is speaking a short phrase, the raw dynamic face

sequence contains a 2D intensity sequence and a pixel-wise registered 3D point cloud sequence. From the 2D intensity sequence, 2D facial landmark (FLM) detection and tracking [45] are performed to obtain consecutive 2D FLMs. More details can be seen in Fig.4. Meanwhile, from the corresponding 3D point cloud sequence, 4D spatio-temporal fusion guided by 2D intensity tracking [46] is performed to reduce 3D spatial noise and temporal fluctuations. Because the 2D and 3D images are registered, the 2D FLMs also specify the corresponding 3D FLMs. i.e., using the 2D FLM sequence. Both a pixel-wise registered 3D FLM sequence and 3D mesh patches constructed around each 3D FLM can be extracted from the denoised 3D sequence.

Subsequently, we construct keypoint-based signatures using pre-defined inter-FLM distances and local-shape-based signatures using the 3D mesh patches. More details are given in Table 3 and Section 4.2. The aggregation of statistics of the signatures gives a compact feature vector that encodes both local geometrical shape information and topological structure information of the 3D dynamic face. Then, all the components in the feature vector are normalized using a multi-dimensional Gaussian model fitted to the whole feature space, resulting in a scale-invariant 3D dynamic face feature descriptor. To reduce the influence of noisy components and improve learning efficiency, discriminative components in the full feature descriptor are selected according to a feature separability metric.

Using the proposed 3D facial dynamics representation, each speaker's 3D facial motion is encoded as a feature descriptor in the offline stage. We employ subject-specific linear discriminant analysis to train an ensemble of linear classifiers over the normalized feature space. Note that each classifier in the ensemble can be regarded as a high-level 3D facial dynamics representation of the subject. We collect all the ensembles in a pre-trained database. In the online stage, a test probe represented by its feature descriptor is compared to the database. The classification with the highest confidence score determines the identity of the tested speaker. We now explain the approach in more detail.

4.2 Speech-driven 3D facial dynamics representation

The proposed hierarchical representation of speech-driven 3D facial dynamics consists of two levels: (1) mid-level representation with part-based 3D dynamic facial primitives; (2) high-level representation with 3D discriminative facial features.

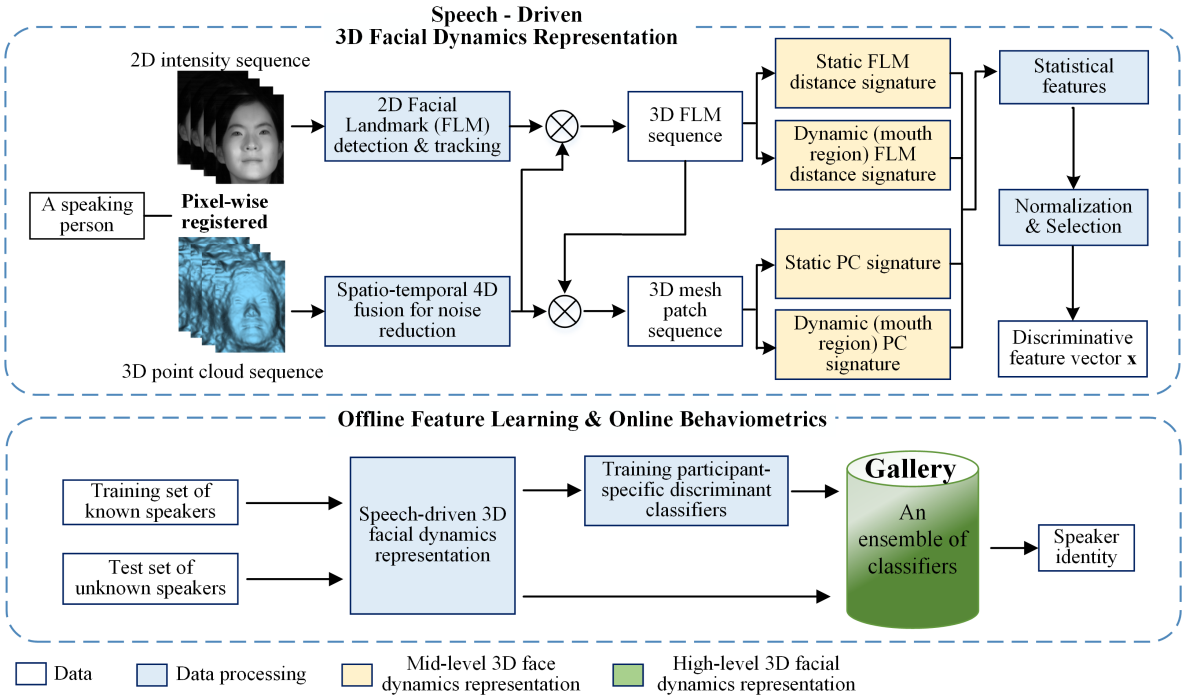

Figure 3: System framework of the proposed behaviometrics. The top box shows the process for extracting the descriptive features from the 2D/3D video. FLMs are firstly detected in the 2D video frames, and then the 3D position of these FLMs are extracted from the registered 3D point cloud (after noise reduction based on fusing multiple consecutive frames). The distances between the static and dynamic FLMs, and the principal curvatures at selected FLMs are extracted for each frame and then aggregated over time as meta-signatures. For each meta-signature, statistical values encoding the static or dynamic property of the speaking face are extracted as meta-features. Finally, the aggregated features are normalized and the best performing subset is selected. The bottom box illustrates how the descriptors extracted using the process in the top box are used for classifier training and then recognition.

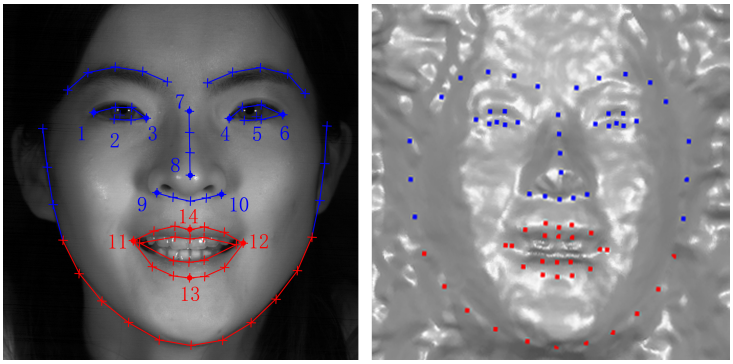

Figure 4: (a) Static facial landmarks (blue) and dynamic facial landmarks (red) on a dynamic face. Number-marked points are facial landmarks of interest; (b) corresponding 3D facial landmarks.

4.2.1 3D speaking face primitives

A speaking face can be separated into static regions R (eyes & nose) and dynamic regions Q (mouth & chin) using 70 FLMs, as shown in Fig.4. We detect the FLMs using an ensemble of regression trees [45]. For each 3D frame t, we select 10 static 3D FLMs of interest $\{\mathbf{p}_n^t \in \mathcal{R}^3\}(n = 1 \cdots 10)$ from the region R and 4 dynamic 3D FLMs $\{\mathbf{q}_m^t \in \mathcal{R}^3\}(m = 1 \cdots 4)$ from the region Q, as the numbers marked in Fig.4. The selected 3D FLMs are used for constructing 10 dynamic and 9 static 3D facial primitives of the speaking face. The meanings and mathematical labels of the defined 3D facial primitives are listed in Table 3.

For each facial primitive, all the samples across a whole sequence form a facial signature. The speaking face results in 4 kinds of signatures, including 5 static FLM distance signatures $\mathbf{SD}_a = \{SD_a^t\}(a = 1 \cdots 5)$, 2 dynamic FLM distance signatures $\mathbf{DD}_b = \{DD_b^t\}(b = 1, 2)$, 4 static Principal Curvature (PC) signatures $\mathbf{SC}_c = \{SC_c^t\}(c = 1 \cdots 4)$, and 8 dynamic PC signatures $\mathbf{DC}_d = \{DC_d^t\}(d = 1 \cdots 8)$. The PCs of a FLM are computed from the neighboring 3D mesh patch $S^t(u, v)$ (with general parameters u and v) of the FLM. As the part-based facial primitives/signatures only represent local properties of the speaking face, we regard them as the mid-level representation of the 3D dynamic face.

We calculate a summary statistical feature from each facial signature. For a static facial signature $\mathbf{SD}_a$ or $\mathbf{SC}_c$, the feature is $f_0(\mathbf{x})$. For a dynamic

facial signature $\mathbf{DD}_b$ or $\mathbf{DC}_d$, the feature is $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, as

$$feature = \begin{cases} f_0(\mathbf{x}) = \frac{1}{n}\sum_t x^t & if \quad \mathbf{x} = \mathbf{SD}_a or \mathbf{SC}_c \\ f_1(\mathbf{x}) = \max_t(x^t) - \frac{1}{n}\sum_t x^t & \\ f_2(\mathbf{x}) = \frac{1}{n}\sum_t x^t - \min_t(x^t) & if \quad \mathbf{x} = \mathbf{DD}_b or \mathbf{DC}_d \end{cases} \tag{1}$$

Combining all the statistical features, a 3D dynamic face is then represented as a 29D feature vector:

$$\mathbf{f} = \begin{array}{l} [w_a f_0(\mathbf{SD}_a), w_b f_2(\mathbf{DD}_b), w_b f_1(\mathbf{DD}_b), \\ w_c f_0(\mathbf{SC}_c), w_d f_1(\mathbf{DC}_d), w_d f_2(\mathbf{DC}_d)] \end{array} \tag{2}$$

where $w_a(a = 1 \cdots 5)$, $w_b(b = 1, 2)$, $w_c(c = 1 \cdots 4)$, $w_d(d = 1 \cdots 8)$ are binary weights assigned to each feature candidate. The weights are determined by discriminative feature selection (detailed in the next section).

4.2.2 Learned representation of 3D speaking face

The values of the raw features are normalized for use during feature selection, training, and classification. The feature space $\{\mathbf{f}\}$ of the 4D face features over all training samples is fitted to a Gaussian model with the mean μ and the covariance $\mathbf{\Sigma}$. The raw feature vector $\mathbf{f}$ is normalized and decorrelated using a whitening transformation:

$$\mathbf{g} = (\mathbf{f} - \mu)\mathbf{\Sigma}^{-1/2} \tag{3}$$

All the raw features are transformed into the whitened space.

To avoid noisy and less-distinctive features degrading the stability and discriminability of the 3D facial dynamics representation, a discriminative feature selection strategy over the normalized feature vector $\{\mathbf{g}\}$ is used to remove less useful feature components. Since the speech-driven 3D facial motion is unique for each subject, we perform discriminative feature selection individually for each subject, producing a subject-specific discriminative 3D facial dynamics representation.

In detail, we measure the separability of the subjects represented as full feature descriptors using the ratio of intra-subject scatter and inter-subject scatter. In the selection process, we iteratively search for the feature components among the full feature vector, resulting in the smallest intra-subject scatter and the largest inter-subject scatter at the same time. The selected

Table 3: Speech-driven 3D dynamic face primitives

10 dynamic primitives in frame t		
Mouth width	DD_1^t	Distance between FLM 11 and 12
Mouth opening	DD_2^t	Distance between FLM 13 and 14
Left mouth corner	DC_1^t, DC_2^t	Max and Min PC of FLM 11
Right mouth corner	DC_3^t, DC_4^t	Max and Min PC of FLM 12
Upper lip	DC_5^t, DC_6^t	Max and Min PC of FLM 13
Lower lip	DC_7^t, DC_8^t	Max and Min PC of FLM 14
9 static primitives in frame t		
Left eye width	SD_1^t	Distance between FLM 1 and 3
Right eye width	SD_2^t	Distance between FLM 4 and 6
L-R eye separation	SD_3^t	Distance between FLM 2 and 5
Nose length	SD_4^t	Distance between FLM 7 and 8
Nose width	SD_5^t	Distance between FLM 9 and 10
Nose bridge	SC_1^t, SC_2^t	Max and Min PC of FLM 7
Nose tip	SC_3^t, SC_4^t	Max and Min PC of FLM 8

FLM: Facial Landmark; PC: Principal Curvature.

(See Fig.4 for FLM locations)

feature components are regarded as the most discriminative from the full feature descriptor for characterizing the subject. Mathematically, we assigned a binary weight vector $\mathbf{w}_s \in \{0,1\}^{29}$ to each subject s, which minimizes the objective function as

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \sum_s \frac{\sum_{\mathbf{g}_s} (\mathbf{g}_s - \mu_s)(\mathbf{g}_s - \mu_s)^{\mathrm{T}}}{(\mu_s - \mu)(\mu_s - \mu)^{\mathrm{T}}} \tag{4}$$

$\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_s]$ aggregates all the subject-specific binary weight vectors $\mathbf{w}_s = [w_1, w_2, \cdots w_{29}]$, $\mathbf{g_s} = [w_1 g_1, w_2 g_2, \cdots w_{29} g_{29}]$ is the feature filtered with the binary weights, $\mu_s \in \mathcal{R}^{29}$ is the weighted mean of all the filtered samples of subject s, and here $\mu \in \mathcal{R}^{29}$ is the weighted mean of all the filtered samples in the transformed feature space $\{\mathbf{g}\}$. The numerator term represents the

intra-subject scatter using the variance of all the samples of the subject s. i.e., the intra-subject scatter measures the difference of feature descriptors of the subject s. The denominator term represents the inter-subject scatter using the Euclidean distance of the subject means. The objective function of this type of feature selection has a correlation with the mechanism of linear discriminant analysis (LDA), while the feature space doesn't change as in LDA. The discriminative feature selection results in a more compact feature descriptor $\mathbf{g}_s \in \mathcal{R}^n$, where n is the number of the selected features with $w_i = 1$ for each subject.

In the training stage, we train an ensemble of subject-specific LDAs using n-fold cross validation. For each subject s, a linear discriminant classifier (LDA classifier) parameterized as $\mathbf{L}_s \in \mathcal{R}^n, b_s \in \mathcal{R}$ is trained using the training samples from the subject s as positive exemplars and the training samples of the remaining subjects as negatives. The subject-specific classification strategy emphasizes discrimination of each subject class from the competing classes and thus increases the recognition power. It also has a similarity to the exemplar-SVM [47], although the E-SVM focuses on intra-class details via an ensemble classifier. The trained model in our classifier is

$$y = \mathbf{L}_s \mathbf{w}_s \mathbf{g} + b_s \tag{5}$$

where $\mathbf{g}$ is the discriminative feature descriptor of a training sample.

In the speaker recognition phase, given a probe sample $\mathbf{g}$, we get the response scores of all the candidate classifiers in the pre-stored dataset. The class with the highest response score $\hat{y}$ is selected as the identity label of the tested participant.

The recognition strategy we use here is consistent with the feature selection process in terms of the objective mechanism. The parameters of each classifier are also a learned representation specific to each subject. Compared with lower-level representations, the learned representation has discriminant capability with holistic semantic information. Thus, we regard it as the high-level representation of a 3D speaking face. Overall, each subject s is represented using a mid-level semantic feature descriptor $\mathbf{f_s}$, a discriminability index $\mathbf{w_s}$ and a high-level holistic feature descriptor $(\mathbf{L_s}, b_s)$.

5 Results and Discussion

The proposed 3D behaviometrics algorithm was verified on the new database S3DFM. We first analyze multiple properties of the 3D speaking face signatures (Sec.5.1) and its derived 3D speaking face features (Sec.5.2). The performance of different 3D face features are compared in Sec.5.3. For reference, we give the performance of 2D face recognition via deep neural networks (DNN) (Sec.5.4) and "shallow" features (Sec.5.5), using 2D intensity samples from the new dataset. We also present discussion on anti-spoofing, head pose variation, video frame rates, and applicability.

5.1 3D speaking face signatures

We investigate the distinctiveness and repeatability of the 3D face primitives defined in Table 3. For each sequence, we extracted static and dynamic face signatures and fitted a line or a curve for each signature.

Examples of static and dynamic face signatures of one subject (Fig.4) are shown in Fig.5. Qualitatively, the defined static facial signatures (SD_i, SC_i) are stable across the sequence. The dynamic facial signatures are distinctive for describing the speech-driven 3D facial motion. Quantitatively, it is supposed that the nose bridge (SC_1, SC_2) and tip (SC_3, SC_4) are the most discriminating static facial primitives in terms of principle curvatures. The standard deviations of the nose bridge SC_1 and tip SC_3 are 11.33 and 1.62, respectively. If we assume that the measure range is between -150 and 200, the relative standard deviations of the signatures are 3.23% and 0.46%, respectively, which demonstrates that the static PC-based facial signatures are also stable.

Fig.6 shows the repeatability and distinctiveness of the mouth-related signatures DD_1, DD_2. The mouth width DD_1 and opening DD_2 are correlated, thus changing simultaneously when the subject is speaking. The fitted curves in Fig.6b and 6d show the movement pattern clearly. Fig.6e and 6f compares 5 example mouth width DD_1 and mouth opening DD_2 from 5 different subjects, which obviously shows the distinctiveness of the two dynamic facial signatures across different subjects.

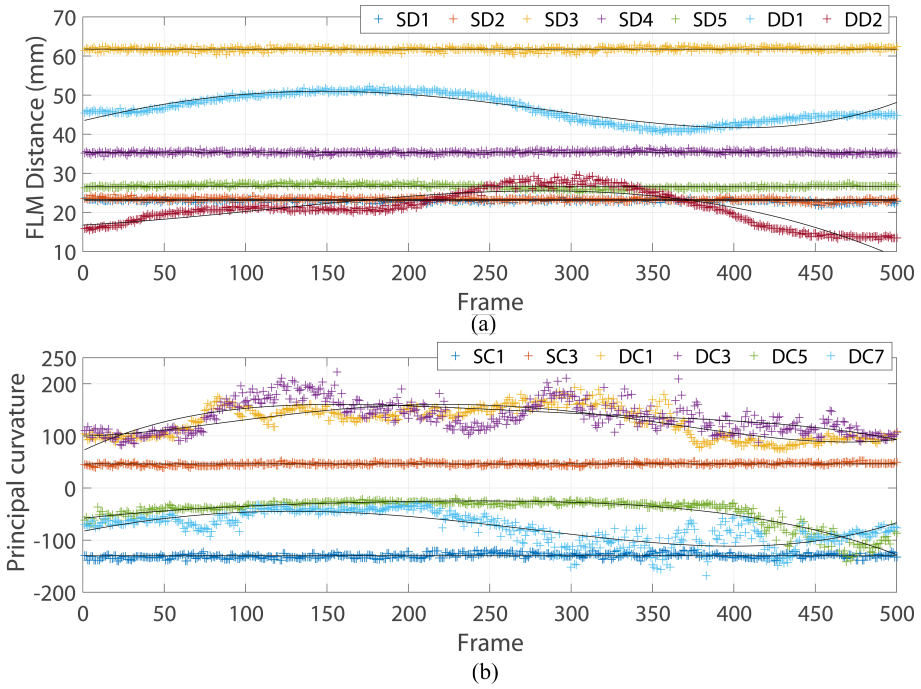

Figure 5: (a) 7 FLM-distance-based facial signatures; (b) 6 PC-based facial signatures. The solid lines are the best line or q-spline fits.

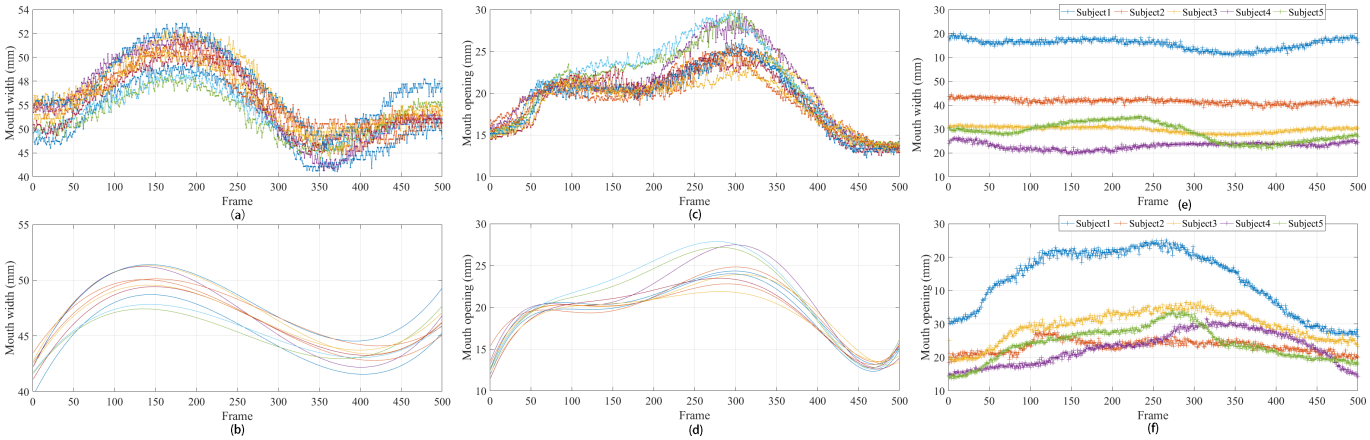

Figure 6: Repeatability & Distinctiveness: (a) 10 instances of mouth width signature (DD_1) from one subject; (b) 10 instances of mouth opening signature (DD_2) from the same subject; (c) 10 fitted mouth width signatures from the same subject; (d) 10 fitted mouth opening signatures from the same subject; (e) 5 instances of mouth width signature (DD_1) from 5 subjects; (f) 5 instances of mouth opening signature (DD_2) from the same 5 subjects.

5.2 Discriminative 3D speaking face features

5.2.1 Similarity evaluation of full features

To quantitatively evaluate the similarities of the features and subjects, we measured the similarity distances between full feature vectors of the 3D facial motion using the Mahalanobis distance. Firstly, each participant was regarded as an independent class and each sequence was represented as a 29D feature vector. We calculated the mean Mahalanobis distance between each pair of classes. Specifically, for each pair of participants {A, B}, we computed the average distance between each of the 10 samples from A and each of the 10 samples from B. The similarity matrix of the 77 classes is shown in Fig.7a. In addition, regarding the 29 features as 29 independent classes, we investigate the correlation of each pair of the 29 features over the 770 samples. The correlation matrix of the 29 features is shown in Fig.7b.

In Fig.7a, 76 Mahalanobis distances along the diagonal have the minimum value along the same row or column, which suggests that, on average, 98.7% of the 77 subjects are well represented by the proposed feature set. Fig.7b shows that the diagonal values are maximum among the correlation

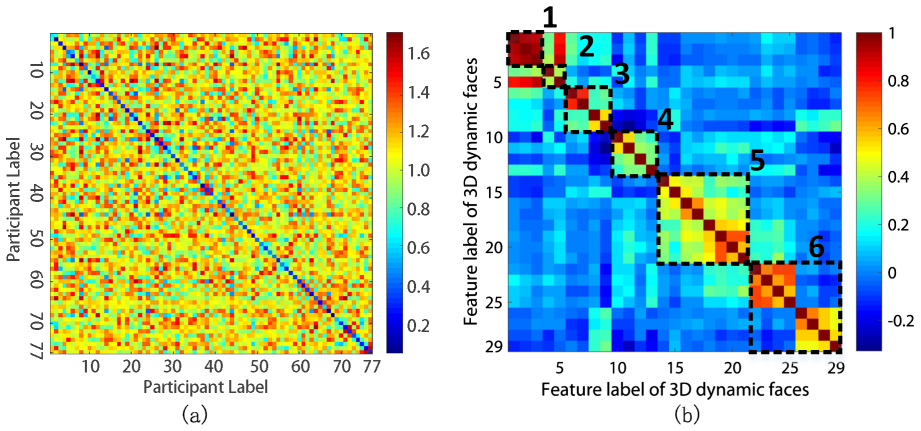

Figure 7: (a) Similarity matrix of different subjects (blue-ish colors indicate greater similarity); (b) Correlation matrix of different features (blue-ish colors indicate less correlation). The features are grouped by regions. Region 1 is static FLM distances of eyes; region 2 is static FLM distances of noses; region 3 is dynamic FLM distances of mouth; region 4 is static PCs of noses; region 5 is dynamic PCs of mouth corners; region 6 is dynamic PCs of lips.

values in the same row, which means that all the 29 features are effective for representing the 3D facial motion. However, the features extracted from the same face region are relatively highly correlated with each other, especially for the FLM-based features of eyes (region 1). Therefore, discriminative feature selection is used for removing redundant features.

5.2.2 Selected feature evaluation

Discriminative feature selection based on the Sequential Forward Selection strategy [48] was conducted in an iterative fashion over the 770 samples. The most useful features are sorted out from all the features one by one. For each round, one feature is sorted out, which minimizes the mean recognition error of 5-fold cross validation. The selection procedure sequentially continues until the mean recognition error does not decrease any more. The results of the sequential selection process are shown in Fig.8, and the corresponding semantic information of the selected features is listed in table 4. From Fig.8a, the 5-fold cross validation error gradually converges to a minimum value after 18 features are selected. The first 10 features contain 5 static features and 5 dynamic features, which means that both static and dynamic features

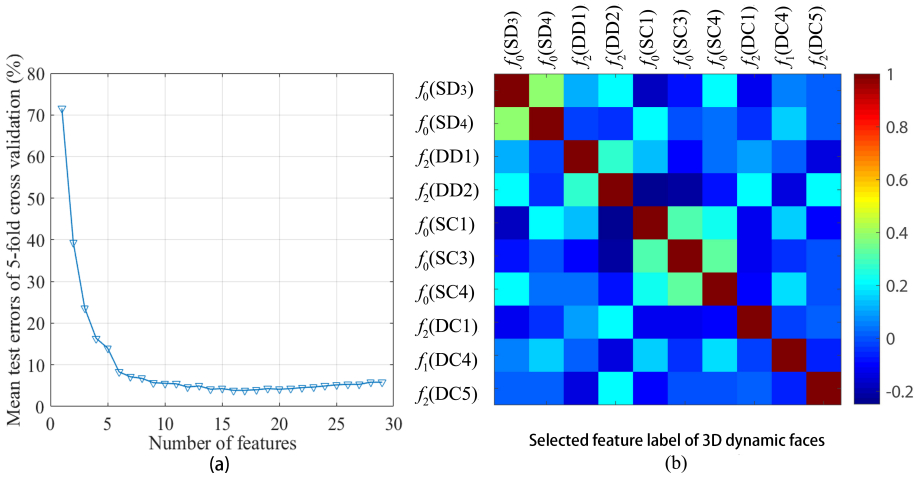

Figure 8: (a) Selected features vs. average 5-fold cross validation accuracy; (b) correlation matrix of selected features (refer to Eqn.(1) for the mathematical symbols).

Table 4: Sequentially selected features

Order	**Feature Meaning**	**Symbol**
1	L-R eye separation	$f_0(SD_3)$
2	Nose length	$f_0(SD_4)$
3	Max PC of nose tip	$f_0(SC_3)$
4	Min PC of nose tip	$f_0(SC_4)$
5	Mouth opening	$f_2(DD_2)$
6	Max PC of nose bridge	$f_0(SC_1)$
7	Mouth width	$f_2(DD_1)$
8	Max PC of left mouth corner	$f_2(DC_1)$
9	Max PC of upper lip	$f_2(DC_5)$
10	Min PC of right mouth corner	$f_1(DC_4)$

For the mathematical symbols, refer to Eqn.(1).

contribute to the behaviometrics performance. Fig.8b shows the similarity matrix of the top 10 features. Compared with the full features in Fig.7b, the selected features have less correlation and higher distinctiveness.

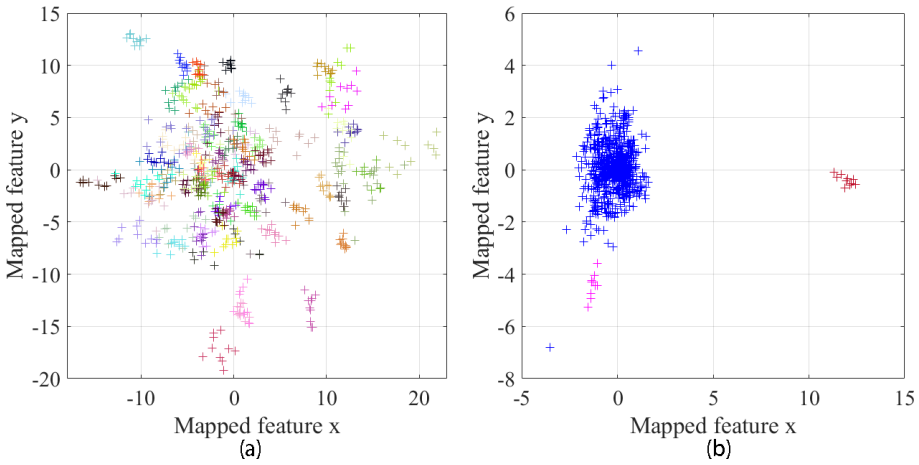

Figure 9: Distribution of dimensionality-reduced features: (a) using multi-class LDA, where each color is a different subject and good clusters can be seen; (b) using subject-specific LDA, where the dark red and rose red samples are from 2 subjects and the blue samples are from the remaining subjects.

5.2.3 Distribution of 3D speaking face features

To investigate separability of all the samples, we trained a generic LDA classifier and mapped all the samples represented by the full 29D features to a 2D feature space. The mapped 2D features produced by the generic LDA classifier can be seen in Fig.9a, where each subject (with 10 samples) has its own color. We also relabeled all the samples into three classes: a pair of subjects become the first two classes, and the remaining samples are the third class. A specific LDA classifier was trained over the three classes, resulting in a distribution of the mapped features as shown in Fig.9b.

Overall, all the samples in Fig.9a cluster well, suggesting that (1) the passcode descriptor is repeatable and robust for each individual; (2) individuals are largely distinguishable. However, as the number of classes gets larger and larger, the overlapping region of the classes would expand as well, which decreases the separability of the samples. Compared with Fig.9a, the mapped features (in Fig.9b) reduced by a specific LDA are more discriminative and separable from negatives, even for the classes whose repeatability is not high.

Table 5: Recognition rate with different features from speech-driven 3D facial motion (%)

	Proposed	Linear SVM	Mahal KNN (k=3)	Simple Trees	Bagged Trees	Mean results
3D Static Features	86.9 ± 0.49	71.0 ± 1.05	85.7 ± 0.51	83.5 ± 0.99	93.8 ± 0.67	84.2 ± 0.74
3D Dynamic Features	85.6 ± 0.53	72.1 ± 0.76	74.3 ± 0.70	76.2 ± 1.13	85.5 ± 0.93	78.7 ± 0.81
Full Features	93.6 ± 0.34	93.6 ± 0.46	87.4 ± 0.58	89.6 ± 1.13	91.9 ± 0.78	91.2 ± 0.66
3D Selected Features (Proposed)	$\mathbf{96.1 \pm 0.28}$	89.0 ± 0.55	93.4 ± 0.48	89.9 ± 0.87	93.0 ± 0.69	92.3 ± 0.57

5.3 Recognition performance on different 3D feature sets

To examine the discriminating capability of the selected features, we verified the proposed 3D behaviometrics algorithm in comparison to the results from 4 other basic classifier pipelines (Linear SVM, Mahal KNN, Simple Trees, Bagged Trees) versus four 3D feature sets (3D Static Feature, 3D Dynamic Feature, 3D Full Feature, 3D Selected Feature). The single feature sets were constructed using the subsets of the 29 features. The results are listed in Table 5. It is obvious that the 3D-Selected-Feature-based pipelines achieve superior performance under the same classifier. The comparison of 3D-Static-Feature-based algorithms and 3D-Selected-Feature-based algorithms illustrates that the discriminative static and dynamic features improve the speaker identification performance. Among the pipelines based on 3D Selected Features, our proposed pipeline outperforms the others, with the highest mean recognition rate of 96.1%.

We also investigated the Cumulative Match Characteristic (CMC) curves of different 3D features sets and 2D-3D joint features using the proposed pipeline. The results are presented in Fig.10. The rank-N responses were ordered by matching scores. Fig.10a demonstrates conclusions consistent with Section 5.5 (i.e., adding 3D features and feature selection boosts performance). Fig.10b shows that the Selected 3D Features perform best in terms of the rank-1 recognition rate, followed by the Full 3D Features. However, the 3D PC-based features (either static or dynamic) perform poorly across the ranks. That is because the principal curvatures are more sensitive to 3D noise, which decreases the stability of 3D PC-based features.

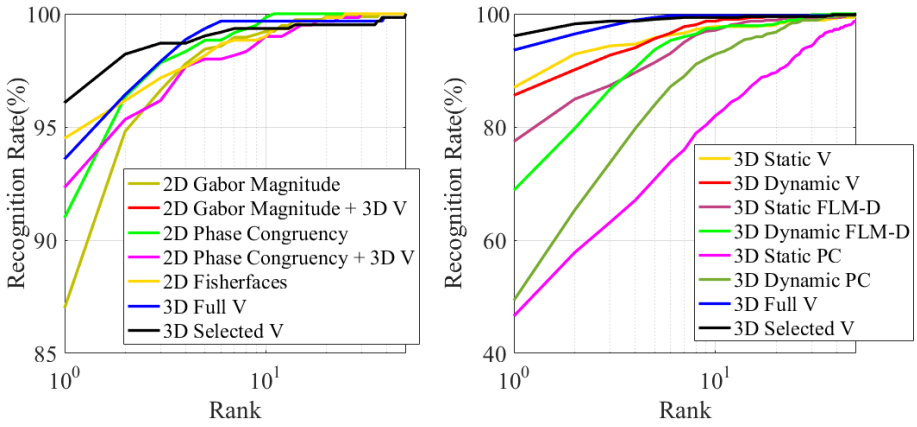

Figure 10: (a) CMC curves of 2D & 2D-3D joint feature sets using the proposed pipeline; (b) CMC curves of different 3D feature sets using the proposed pipeline. (V is the simplification of Visual; FLM-D is FLM distance; PC is principal curvature.)

5.4 Discussion on deep-neural-network based 2D face biometrics and the proposed behaviometrics

In this section, we show the performance of two popular DNNs [49, 50] specific for face recognition, using the 2D intensity samples from the new dataset. A performance-based comparison of 2D face biometrics and 3D face behaviometrics is not completely ideal because of the properties of different modalities (discussed after the performance test). Instead, we analyze that both 2D face and 3D face behaviometric algorithms can be applied in different conditions.

The top performing 2D face recognition algorithms used different high-dimensional DNN-based face descriptors and achieved remarkable performance over a large dataset "Labeled Faces in the Wild" (LFW). For example, FaceNet [49] achieves 99.63% verification accuracy; DeepID 3 [51] achieves 99.53% verification accuracy and 96.0% identification rate; DeepFace [52] has 97.35% verification accuracy; VGG-Face [50] has 97.27% verification accuracy. We tested FaceNet and VGG-Face on the 2D samples (77 subjects) from the S3DFM dataset. For each subject, we randomly sampled 4 static frames per sequence (4 frames $\times$ 10 sequences per subject, 40 frames $\times$ 77 subjects in total). Each DNN model was initialized using a pre-trained model, then it is adjusted and fine-tuned on the new dataset via 5-fold cross validation.

Table 6: Performance of DNN-based 2D face recognition algorithms on our 2D intensity samples(%)

DNN model	Input frame (pixels)	Feature number	Recognition rate (%)
FaceNet	224×224	4096	79.2
VGG-Face	600×600	128	99.4

The input frames are without alignment & cropped

The overall performance is the mean recognition rate of test samples over the 5 folds. The feature dimensionality and performance of the two DNNs on our 2D samples are listed in Table 6. In comparison with the benchmark results of the DNN models, our test was on a relatively small scale dataset. We did not pre-process images with face alignment and detection due to the pure background of the images.

Although many 2D face biometric algorithms have been developed, the 3D face dynamics approach we propose here has some extra properties, in addition to its good performance. While not quite as good as VGG-Face on our dataset, the 3D face dynamics approach gives a private passcode that has double value, namely: 1) the individual's choice of the passphrase, and 2) the subject-specific bio-modality arising from their individual mouth motions. The behavior is performed highly collaboratively and is hard to be captured unconsciously and reproduced elsewhere. These properties make the 3D behavior hard to imitate, e.g. for criminal purposes. Thus, our pipeline is suitable to be applications that require a higher anti-spoofing level.

2D face data is easily captured and provides rich texture information, but it lacks real 3D geometric information due to the image projection. The challenges in 2D face recognition always result from sensitivity to face orientation, scale, appearance, lighting, etc. Deep representations have been striving to minimize intra-class variations, but they need large training datasets. Compared with 2D face recognition, a limitation of the 3D face community is that the size of existing 3D datasets is smaller.

Table 7: Performance of 2D & 2D+3D face recognition on the new dataset

Algorithm	Number of features	Recognition rate (%)
2D Gabor Magnitude + PCA	29	87.0 ± 1.03
2D Phase Congruency + PCA	29	95.2 ± 0.34
2D Fisherface + PCA	29	94.5 ± 0.46
2D Gabor Magitude + 3D Visual	58	91.0 ± 0.69
2D Phase Congruency + 3D Visual	58	92.3 ± 0.58
3D Visual	29	93.6 ± 0.34
Selected 3D Visual (Ours)	18	$\mathbf{96.1 \pm 0.28}$

5.5 "Shallow"-feature-based 2D face recognition on the intensity samples of S3DFM

We investigated the performance of 3 well-established 2D face recognition pipelines via algorithmic "shallow" features extracted from our 2D intensity samples, and investigate the performance of 2D-3D combined face modalities at person identification. The algorithmic "shallow" features are constructed with explicit modelling or structure, in contrast to the deep-net-based features in section 5.4. The compared features here are Phase Congruency features [53], Gabor Magnitude features [54] and Fisherfaces [55].

The 2D face identification experiment was conducted using the same 2D intensity dataset (with 3080 samples) as above. The training and identification used 5-fold cross validation. For each trial, we regarded one split as test samples and used the remaining 4 splits for training a classifier. The recognition rate was calculated by counting correctly classified samples across all the 5 trials. The recognition rates of the different pipelines are compared in Table 7. One can see that for the 2D Gabor Magitude, the combined 3D visual features help improve the performance, while for the 2D Phase Congruency, the recognition rate drops slightly. That means the 2D-3D joint features do not always increase the performance, but the discriminative features play the most important role in the task.

5.6 Robustness against varying head poses

We mixed the 770 frontal face samples with 260 samples with continuously changing head pose for data training and test. The head motion includes

rotation, pitching, and yawing. We calculated the changing 3D pose of each 3D face sequence via landmark tracking. Three examples of head motions and an example of a frontal head are shown in Fig.11, where four representative frames with tracked 2D facial landmarks are also shown above the 3D pose curves of each motion. Then, we extracted facial signatures and constructed a statistical feature for each face sequence, using the frames where the yaw and pitch angles are within 10 degrees relative to frontal pose. The training and test were performed through discriminative feature selection, minimizing the mean recognition rate of 5-fold cross validation. The overall recognition error rate when using sequential feature selection converges to a minimum, as shown in Fig.12b. Overall, the algorithm performed on the mixed samples achieves a best recognition rate of 96.0%, using the first selected 16 features. It demonstrates that the proposed algorithm is robust against head pose changing within 10 degrees.

5.7 Robustness against different frame rates

Our dataset records the 2D-3D face sequences with 500 fps. In order to investigate whether the high frame rate sequences are over-sampling (contain redundant frames) for biometrics, we subsampled all the sequences to lower frame rates varying from 25 fps to 500 fps respectively and computed the mean recognition rate of the proposed behaviometric algorithm. The result is shown in Fig.12a. The performance degrades slightly when the frame rate is less than 100 fps, while higher frame rates (125 fps to 500 fps) scarcely improve the performance. It shows that 100 fps video is compatible with the speech-driven facial dynamics. However, we preserve the raw sequences with 500 fps in the public dataset, since the extra frames are also helpful in generating improved 3D data by fusion-based spatio-temporal noise reduction algorithms [56, 57].

5.8 Robustness to spoofing

This section presents qualitative and quantitative analyses on anti-spoofing. Compared with 2D static biometric data, 3D behaviometric data usually has a higher imitation difficulty. The proposed 3D speech-driven face modality includes evidence from person-specific face motion. The 3D dynamics are less likely to be stolen and mimicked, and help guard against occlusion-level spoofing (masks) and appearance-level spoofing (painting, makeup). An im-

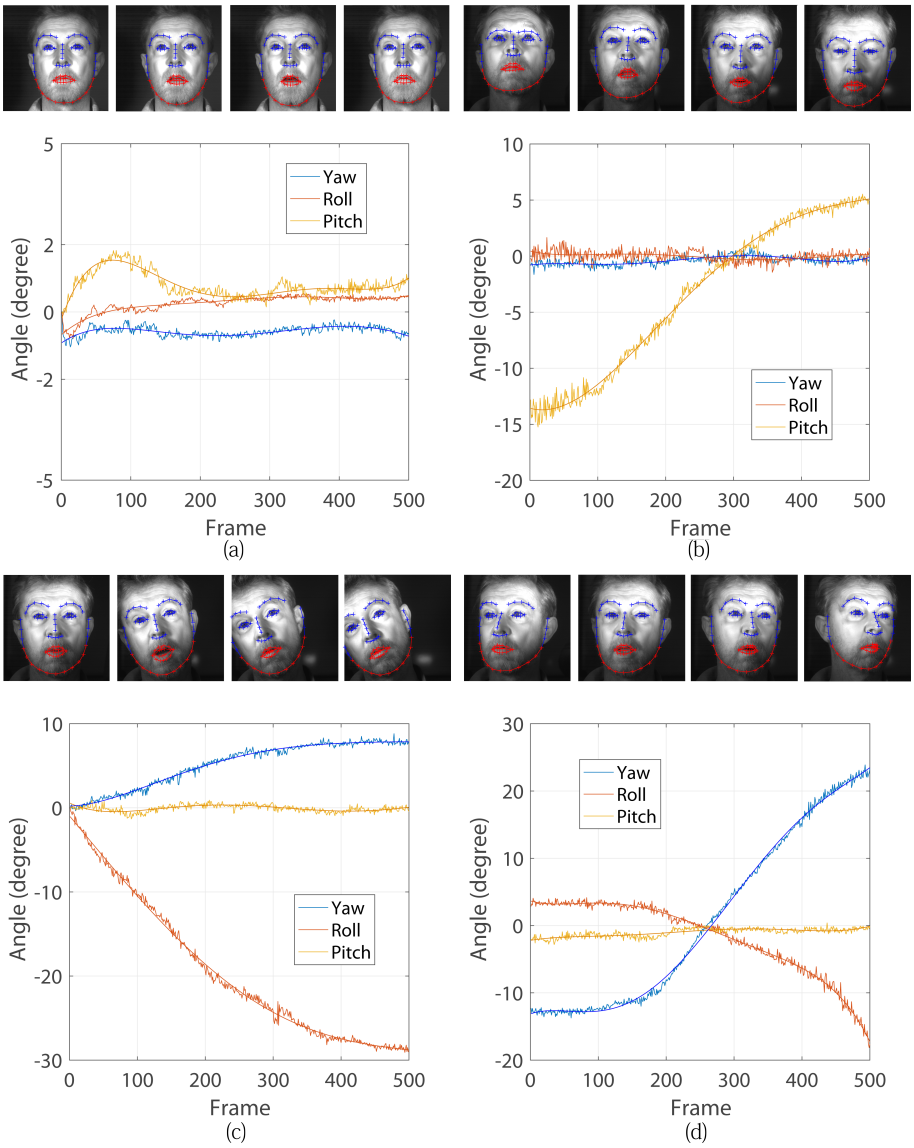

Figure 11: 3D pose estimation of head motions and frontal head: (a) frontal head; (b) pitching; (c) rotating; (d) yawing.

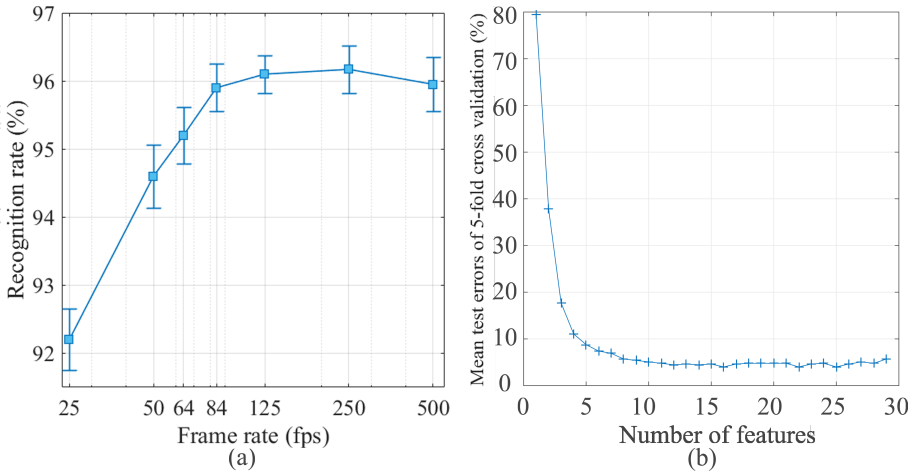

Figure 12: (a) Recognition rate vs. Frame rate; (b) Feature selection during recognition on mixed samples.

postor would need to match the subject-specific speech-driven motion, which is the first-level prerequisite for passing the guard. Overall, the proposed 3D behaviometrics have an intuitive advantage beyond the quantitative result in enhancing information security.

In the S3DFM Dataset, since all the participants use the same passcode, it is exactly the situation of imitation spoofing. We assume that only one participant with the passcode is a genuine client, and all the rest are impostors. We calculated the ROC curve of the proposed pipeline using selected 3D features, and compared it with those of using 3D static or 3D dynamic features. The results are shown in Fig.13. In Fig.13a, the selected 3D features demonstrate the strongest separability when measuring client and spoofing scores, with 0 EER. Using the pure static features and dynamic features by themselves show an EER of 11.53% and 8.90%, respectively.

5.9 Discussion on applicability

The applicability of the proposed pipeline is discussed from 4 aspects: main use case, passcode, speaking speed, and computational cost.

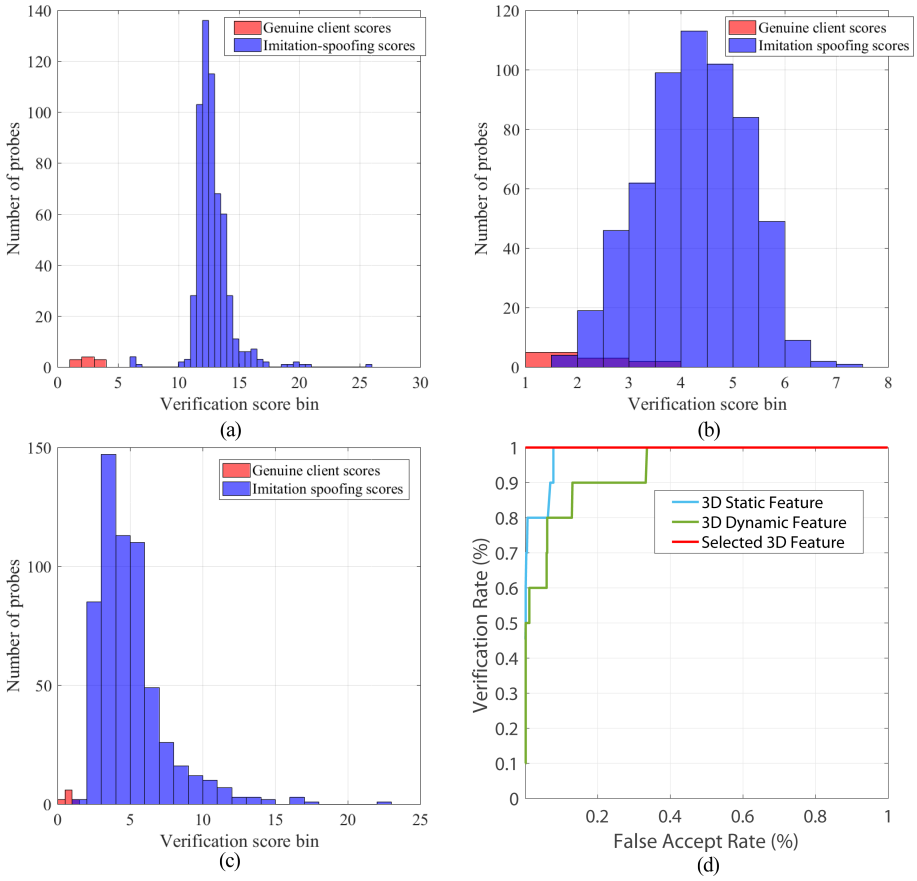

Figure 13: Verification score distribution and ROC curve: (a) selected 3D features; (b) 3D static features; (c) 3D dynamic features; (d) ROC curves.

5.9.1 Main use case

The proposed algorithm can be applied in any biometric system that aims for high robustness against static face spoofing, such as 2D/3D mask occlusion. For example, it could be deployed at a high-security door entry system, a cash machine or places where fixed 3D video scanners could be installed facing users. To ensure the algorithm works at its best performance, it is suggested that the sensor system is stably installed with low vibration and is calibrated as precisely as possible. The speaking face should be naturally posing frontal to the scanners or is allowed to be moving slightly (within 10 degrees as discussed in Section 5.6). The speech-driven facial dynamics is a prerequisite to access the protected system. That is, a candidate should show a collaborative action by speaking a private passcode (set by the candidate before) in front of the biometrics system, in order to get access.

5.9.2 Passcode

Although we used the same passcode across all the participants here, the proposed algorithm is applicable to any passcode causing repeatable lip motions. In real applications, a client would have a private passcode. Essentially, the speech-driven facial dynamics is a passcode-guided motion. Both the privacy of the passcode and the subject-specific facial dynamics contribute to the uniqueness of the biological modality. It is notable that the constrained behavior (by passcodes) does not degrade the generalization of a biometric system, but increases the spoofing difficulty of bio-modality and the security level of the biometric system instead.

5.9.3 Speaking speed

The feature representation and behaviometrics proposed here are invariant to speaking speed, but we think that speaking speed could be additional bio-information used for biometrics in the future.

5.9.4 Computational cost

The experiments were conducted using a Matlab implementation on a computer with a 3.40 GHz 6-core CPU. The most time-consuming part is training classifiers for constructing an offline gallery, which can be done once in an offline stage and thus does not influence online probe tests. In the online stage,

given a 2D-3D sequence probe of 500 fps, facial landmark tracking across all the frames of 600×600 pixels takes $150.4s$. Actually, for online use, the facial landmark tracking could be replaced by a detection algorithm, which allows the facial landmark extraction to be performed in parallel. The facial landmark detection in one frame takes $0.14s$. Then, the feature construction and classification for a sequence could cost on average $0.12s$. Therefore, with parallel computation, the online test of a probe could be done within $1s$.

6 Conclusions and Future Work

This paper presents a 3D behaviometric pipeline based on speech-driven 3D facial dynamics as a "3D visual passcode" and releases the first publicly available Speech-driven 3D Facial Motion dataset (S3DFM). Experiments on the new dataset verify that (1) the speech-driven dynamic face signatures are repeatable and distinctive. The 29 statistical features have 100% separability (see Fig.7b); the 77 subjects represented by the full 29 features have 98.7% separability (see Fig.7a); (2) The proposed approach improves the identity recognition performance across 5 different classifiers, with the best recognition rate of 96.1%. The comparable result with other feature sets demonstrates the effectiveness of adding dynamic information to the static 3D descriptors; (3) The algorithm is robust against random head movement within 10 degrees of yaw and pitch. The overall recognition rate on mixed samples of frontal and non-frontal talking faces is 96.0%, which means that adding non-frontal face samples hardly degrades the recognition performance.

In the future, we would like to further expand our dataset by adding samples from more participants. The work presented here uses the same passcode for every subject. Clearly, including more distinct passphrases would increase discrimination between people. There are at least 3 major directions for further investigations: (1) The passcode used across the participants is a 2-syllable word. It would be interesting to investigate more types of passphrases with 1, 3 or 4 syllables. The passphrases with good performance could constitute a recommended passcode gallery for users; (2) Temporal modelling or prediction could be merged with the spatial features for stronger speech-driven dynamic face representation; (3) At least 2 additional sources of information might be used for better anti-spoofing: time-sensitive features related with the speaking speed and the synchronization between the 3D video and audio properties.

Acknowledgment

The work was supported by the funding from the China Scholarship Council (CSC) under grant 201606020087. We would like to thank all the participants for their contribution to the database.

References

[1] Y. Dong, H. Wu, X. Li, C. Zhou, and Q. Wu, "Multiscale symmetric dense micro-block difference for texture classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 10.1109/TCSVT.2018.2883825, 2018.

[2] X. Li, G. Cui, and Y. Dong, "Graph regularized non-negative low-rank matrix factorization for image clustering," *IEEE Trans. Cybernetics*, vol. 47, no. 11, pp. 3840–3853, 2017.

[3] Y. Chang, M. Vieira, M. Turk, and L. Velho, "Automatic 3d facial expression analysis in videos," *Analysis and Modelling of Faces and Gestures*, pp. 293–307, 2005.

[4] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3d dynamic facial expression database," in *Automatic Face & Gesture Recognition. 8th IEEE Int. Conf. on*, 2008, pp. 1–6.

[5] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014.

[6] S. Aly, A. Trubanova, L. Abbott, S. White, and A. Youssef, "Vt-kfer: A kinect-based rgbd+ time dataset for spontaneous and non-spontaneous facial expression recognition," in *Biometrics (ICB), 2015 Int. Conf. on*, 2015, pp. 90–97.

[7] L. Benedikt, V. Kajic, D. Cosker, P. L. Rosin, and A. D. Marshall, "Assessing the uniqueness and permanence of facial actions for use in biometric applications," *IEEE Trans. on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 40, no. 3, pp. 449–460, 2010.

[8] B. J. Matuszewski, W. Quan, L.-K. Shark, A. S. Mcloughlin, C. E. Lightbody, H. C. Emsley, and C. L. Watkins, "Hi4d-adsip 3-d dynamic facial articulation database," *Image and Vision Computing*, vol. 30, no. 10, pp. 713–727, 2012.

[9] Y. Dong, D. Tao, X. Li, J. Ma, and J. Pu, "Texture classification and retrieval using shearlets and linear regression," *IEEE Trans. Cybernetics*, vol. 45, no. 3, pp. 358–369, 2015.

[10] Y. Lei, Y. Guo, M. Hayat, M. Bennamoun, and X. Zhou, "A two-phase weighted collaborative representation for 3d partial face recognition with single sample," *Pattern Recognition*, vol. 52, pp. 218–237, 2016.

[11] H. Li, D. Huang, J. Morvan, Y. Wang, and L. Chen, "Towards 3d face recognition in the real: A registration-free approach using fine-grained matching of 3d keypoint descriptors," *Int. J. Comput. Vision*, vol. 113, no. 2, pp. 128–142, 2015.

[12] P. Huber, G. Hu, J. R. Tena, P. Mortazavian, W. P. Koppen, W. J. Christmas, M. Rätsch, and J. Kittler, "A multiresolution 3d morphable face model and fitting framework," in *Proc. of the 11th Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016)*, vol. 4, 2016, pp. 79–86.

[13] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *Sixth IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS)*, 2009, pp. 296–301.

[14] D. Kim, M. Hernandez, J. Choi, and G. G. Medioni, "Deep 3d face identification," *CoRR*, vol. abs/1703.10714, 2017.

[15] M. Tistarelli, M. Bicego, and E. Grosso, "Dynamic face recognition: From human to machine vision," *Image and Vision Computing*, vol. 27, no. 3, pp. 222–232, 2009.

[16] A. Hadid, J.-L. Dugelay, and M. Pietikäinen, "On the use of dynamic features in face biometrics: recent advances and challenges," *Signal, Image and Video Processing*, vol. 5, no. 4, pp. 495–506, 2011.

[17] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *IEEE Trans. on Image Processing*, vol. 15, no. 10, pp. 2879–2891, 2006.

[18] D. Dean and S. Sridharan, "Dynamic visual features for audio–visual speaker verification," *Computer Speech & Language*, vol. 24, no. 2, pp. 136–149, 2010.

[19] E. Boutellaa, Z. Boulkenafet, J. Komulainen, and A. Hadid, "Audio-visual synchrony assessment for replay attack detection in talking face biometrics," *Multimed. Tools Appl.*, vol. 75, no. 9, pp. 5329–5343, 2016.

[20] S. Zafeiriou and M. Pantic, "Facial behaviometrics: The case of facial deformation in spontaneous smile/laughter," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conf. on*, 2011, pp. 13–19.

[21] A. Dantcheva and F. Brémond, "Gender estimation based on smile-dynamics," *IEEE Trans. on Information Forensics and Security*, vol. 12, no. 3, pp. 719–729, 2017.

[22] L. Benedikt, D. Cosker, P. L. Rosin, and D. Marshall, "3d facial gestures in biometrics: from feasibility study to application," in *Biometrics: Theory, Applications and Systems, 2nd IEEE Int. Conf. on*, 2008, pp. 1–6.

[23] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3d facial expression recognition: A comprehensive survey," *Image and Vision Computing*, vol. 30, no. 10, pp. 683–697, 2012.

[24] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[25] Y. Sun and L. Yin, "Facial expression recognition based on 3d dynamic range model sequences," *Computer Vision–ECCV 2008*, pp. 58–71, 2008.

[26] F. Tsalakanidou and S. Malassiotis, "Real-time 2d+ 3d facial action and expression recognition," *Pattern Recognition*, vol. 43, no. 5, pp. 1763–1775, 2010.

[27] V. Le, H. Tang, and T. S. Huang, "Expression recognition from 3d dynamic faces using robust spatio-temporal shape features," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE Int. Conf. on*, 2011, pp. 414–421.

[28] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3d facial expression dynamics," *Image and Vision Computing*, vol. 30, no. 10, pp. 762–773, 2012.

[29] T. Fang, X. Zhao, O. Ocegueda, S. K. Shah, and I. A. Kakadiaris, "3d/4d facial expression analysis: An advanced annotated face model approach," *Image and Vision Computing*, vol. 30, no. 10, pp. 738–749, 2012.

[30] B. B. Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava, "4-d facial expression recognition by learning geometric deformations," *IEEE Trans. on Cybernetics*, vol. 44, no. 12, pp. 2443–2457, 2014.

[31] S. Canavan, Y. Sun, X. Zhang, and L. Yin, "A dynamic curvature based approach for facial activity analysis in 3d space," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conf. on*, 2012, pp. 14–19.

[32] M. Reale, X. Zhang, and L. Yin, "Nebula feature: A space-time feature for posed and spontaneous 4d facial behavior analysis," in *Automatic Face and Gesture Recognition, 10th IEEE Int. Conf. and Workshops on*, 2013, pp. 1–8.

[33] M. Xue, A. Mian, W. Liu, and L. Li, "Automatic 4d facial expression recognition using dct features," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conf. on*, 2015, pp. 199–206.

[34] A. Danelakis, T. Theoharis, I. Pratikakis, and P. Perakis, "An effective methodology for dynamic 3d facial expression retrieval," *Pattern Recognition*, vol. 52, pp. 174–185, 2016.

[35] D. Cosker, E. Krumhuber, and A. Hilton, "A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling," in *Computer Vision (ICCV), 2011 IEEE Int. Conf. on*, 2011, pp. 2296–2303.

[36] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Automatic face and gesture recognition, 7th Int. Conf. on*, 2006, pp. 211–216.

[37] R. M. Jiang, A. H. Sadka, and D. Crookes, "Multimodal biometric human recognition for perceptual human-computer interaction," *IEEE Trans. Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 676–681, 2010.

[38] H. Vajaria, T. Islam, P. K. Mohanty, S. Sarkar, R. Sankar, and R. Kasturi, "Evaluation and analysis of a face and voice outdoor multibiometric system," *Pattern Recogn. Lett.*, vol. 28, no. 12, pp. 1572 – 1580, 2007.

[39] "The vidtimit audio-video dataset," http://conradsanderson.id.au/vidtimit/, accessible on Jan 29, 2019.

[40] "Austalk: An audio-visual corpus of australian english," https://austalk.edu.au/, accessible on Jan 29, 2019.

[41] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matejka, J. Cernockỳ, N. Poh, J. Kittler, A. Larcher, C. Levy *et al.*, "Bi-modal person recognition on a mobile phone: using mobile phone data," in *Multimedia and Expo Workshops (ICMEW), IEEE Int Conf on*, 2012, pp. 635–640.

[42] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[43] "Dimensional imaging (di4d)," http://www.di4d.com/, accessible on Jan 29, 2019.

[44] J. Zhang, K. Richmond, and R. B. Fisher, "Dual-modality talking-metrics: 3d visual-audio integrated behaviometric cues from speakers,"

in *24th International Conference on Pattern Recognition (ICPR)*, Aug 2018, pp. 3144–3149.

[45] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[46] J. Zhang, C. Maniatis, L. Horna, and B. R. Fisher, "Dynamic 3d reconstruction improvement via intensity video guided 4d fusion," *Journal of Visual Communication and Image Representation*, vol. 55, pp. 540–547, 2018.

[47] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-svms for object detection and beyond," in *IEEE Int Conf on Computer Vision (ICCV)*, 2011, pp. 89–96.

[48] A. W. Whitney, "A direct method of nonparametric measurement selection," *IEEE Trans. on Computers*, vol. 100, no. 9, pp. 1100–1103, 1971.

[49] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[50] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. of the British Machine Vision Conf.*, 2015, pp. 41.1–41.12.

[51] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *CoRR*, vol. abs/1502.00873, 2015.

[52] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[53] S. Gundimada, V. K. Asari, and N. Gudur, "Face recognition in multisensor images based on a novel modular feature selection technique," *Information Fusion*, vol. 11, no. 2, pp. 124–132, 2010.

[54] V. Štruc and N. Pavešić, "The complete gabor-fisher classifier for robust face recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, p. 31, 2010.

[55] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[56] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proc. of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 343–352.

[57] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi, "3d scanning deformable objects with a single rgbd sensor," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 493–501.